Learning from Learning Curves: Discovering Interpretable Learning Trajectories

Lujie Chen
Carnegie Mellon University
Pittsburgh, PA
lujiec@andrew.cmu.edu

Artur Dubrawski
Carnegie Mellon University
Pittsburgh, PA
awd@cs.cmu.edu

ABSTRACT

We propose a data driven method for decomposing population level learning curve models into mutually exclusive distinctive groups each consisting of similar learning trajectories. We validate this method on six knowledge components from the log data from an online tutoring system ASSISTment. Preliminary analysis reveals interpretable patterns of "skill growth" that correlate with students' performance in the subsequently administered state standardized tests.

CCS Concepts

•**Information systems** → **Data analytics; Clustering;**

Keywords

Learning curves, Student models, Clustering

1. INTRODUCTION

Learning curves are intuitive tools of descriptive analysis often used in learning science to characterize students' growth of skill or knowledge with experience. A population level skill-specific learning curve may be constructed by aggregating all students' performance at a sequence of time-ordered learning opportunities. This construction implicitly assumes the homogeneity of students' learning trajectories, which may not hold in reality due to the disparity either in terms of prior knowledge or individual learning rates.

Much has been explored to cluster students into sub-groups with the goal to improve prediction. For instance [3] proposed K-means and spectral clustering methods using the dynamic features gathered from students' interaction with an intelligent tutor, the clusters are then used to train ensemble models for predicting post-test scores. The focus of this work is to explicitly model students' learning trajectories for the purpose of discovering and understanding the heterogeneous patterns of "skill growth" of the students as well as their implications on their performance in the subsequent tests.

LAK '17 Vancouver, BC Canada

ACM ISBN 978-1-4503-4870-6/17/03.

DOI: http://dx.doi.org/10.1145/3027385.3029449

2. DATA AND METHODS

We used the 'Assistments Math 2004-2005 (912 Students)' data accessed via DataShop [1]. It contains logs of 912 middle school students interacting with ASSISTment, an online tutoring system that provides help or hints at each step of math problem solving. These logs are linked with the post-test scores from MCAS (Massachusetts Comprehensive Assessment System) of the same set of students. We used the student-step roll up data available from DataShop summarizing students' performance at each learning step which was annotated with a version of a Knowledge Component (KC) model comprised of 39 skills.

For each specific KC, we extracted the first-attempt performance (correct or incorrect) for each student, with each response indexed by the opportunity count that starts with 1 and increments by 1 for each additional opportunity that student encounters. For steps annotated with multiple KCs, we assigned the step level performance for each of the KC involved. For this analysis, we selected 6 KCs with at least several hundred students each of whom had practiced at least 10 times on those KCs.

Based on the student-level performance data described above, we then fit a Group Based Model (GBM, [2]) to discover distinct groups of learning trajectories. GBM models a set of learning curves as a mixture of polynomial functions with timestamps (in our case, the opportunity count indices) as the covariates. For a given learning curve of length T, assuming K groups, the likelihood function of the observed trajectory is provided as follows:

$$\sum_{i=1}^{K} \pi_i \prod_{j=1}^{T} P_i(Y_j|X_j) \tag{1}$$

where π_i and P_i are respectively the prior probability of data belonging to a certain group and the likelihood function for group indexed by i. Y_j is the response variable value at step j, in our case it is student's performance in a binary form; and X_j is the covariate, in our case it is the opportunity count (i.e., $X_j = j$). Given the assumed number of distinct groups of learning trajectories (K) and the chosen order of the polynomials, the model estimates K smooth, polynomial in time trajectories one for each of the groups. The form of the likelihood function P_i is determined by the response variable. In our case, since the response variable is binary value of 0 and 1, we used the Bernoulli distribution. The GBM model also estimates the prior distribution of the group membership across data. The model parameters are inferred using Maximum Likelihood method through a nu-

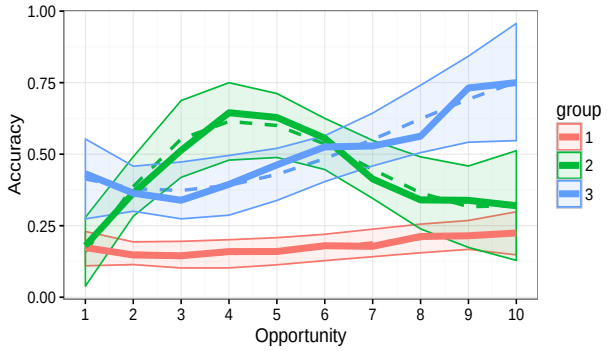

Figure 1: Estimated mean learning curves by group (solid lines) and 95% confidence intervals, overlaid with estimated latent polynomial functions (dotted).

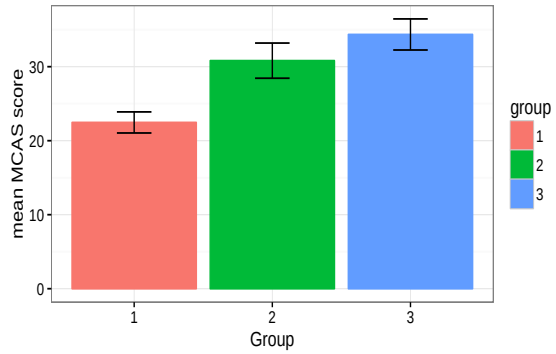

Figure 2: Mean MCAS scores for each of the learning trajectory groups with 95% confidence intervals.

merical optimization procedure. In this experiment, we fit models with $K = 3$ groups and polynomial functions up to the third degree for each group.

Posterior group membership for each individual student's learning curve can then be computed as follows:

$$P(Group = i|Y, X) = \frac{\hat{\pi}_i \prod_{j=1}^{T} \hat{P}_i(Y_j|X_j)}{\sum_{i=1}^{K} \hat{\pi}_i \prod_{j=1}^{T} \hat{P}_i(Y_j|X_j)} \quad (2)$$

In order to assess the predictive utility of discovered groups, we correlate the posterior group membership for each student with their MCAS scores, from there the ANOVA p-values are computed and predictive accuracies are estimated.

3. RESULTS

Figure 1 summarizes one main output from GBM for one of the KC named "P.2.8-evaluating-algebraic-expressions", which is learned by 388 students. The plot shows the learning curve distributions aggregated from students belonging to each of the group as computed from maximum posterior group membership probability 2. The plot is overlaid (dotted line) with model's estimated group trajectories (i.e., the latent polynomial function). As shown, the first group (red) is the most likely to be comprised of students who learned very little within the first 10 steps as suggested by the almost flat learning curve, while the second group (green) represents a subset of students who seemed to learn quickly t first but then slipped into a "forgetful state" as evidenced by the drop of their learning curves after 5 steps. The students in the third group (blue) follow a steadily increasing learning trajectory. Presumably, this type of students are most likely to show good performance in subsequent tests.

We then correlate the posterior group memberships with MCAS scores. As shown in Figure 2, the mean score for the 1st group is the lowest as expected. Students from the 2nd group achieved significantly higher scores on average but not as high as the 3rd group who exhibited "healthy" learning curves. ANOVA p-value less than 0.001 suggests an overall correlation between group membership and post test scores. Similarly significant relationships have been found for all of the other five KCs evaluated.

We further estimated the potential predictive utility of group membership by predicting each student's MCAS test score using the mean score of the group that the student belongs to, based on which we compute Mean Absolute Error (MAE). For the six KCs we studied, MAD ranges from 8.77 to 9.49, which is on par with the performance of the cluster-ensemble model reported in [3].

4. DISCUSSION AND FUTURE WORK

We presented a method to model heterogeneity of students' learning trajectories by employing a group-based approach. In our preliminary analysis that models step-level student log data for 6 different KCs, we noted interesting distinct group level patterns of skill growth that can be readily interpreted. In addition, the observed significant correlation between posterior group membership and MCAS scores suggests that the apparent heterogeneity of learning trajectories is reflected in the students' performance in the future tests. Future work will investigate the likely cause for particular patterns of learning curves (e.g., the forgetting phenomenon in the green plot in Figure 2), to evaluate the opportunity for interventions that might shift students to the more effective trajectories. We will also study the utility of our approach for early detection of group membership when only initial performance data is available, to inform timely interventions.

5. ACKNOWLEDGMENTS

The research reported here was supported, in whole or in part, by the Institute of Education Sciences, U.S.Department of Education, through Grant R305B150008 to Carnegie Mellon University. The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education.

6. REFERENCES

[1] K. R. Koedinger, R. S. J. Baker, K. Cunningham, and A. Skogsholm. A Data Repository for the EDM community : The PSLC DataShop. *Handbook of Educational Data Mining*, pages 43–55, 2010.

[2] D. Nagin. *Group-based modeling of development.* Harvard University Press, Cambridge, MA, 2005.

[3] S. Trivedi, Z. Pardoz, and N. Heffernan. Spectural Clustering in Educational Data MIning. *Proceedings of the 4th International Conference on Educational Data Mining*, pages 129–138, 2011.